ATLAS Open Data For Education

António Jacques Costa, Kate Shaw

LEVERHULME TRUST

MANCHESTER 1824
The University of Manchester

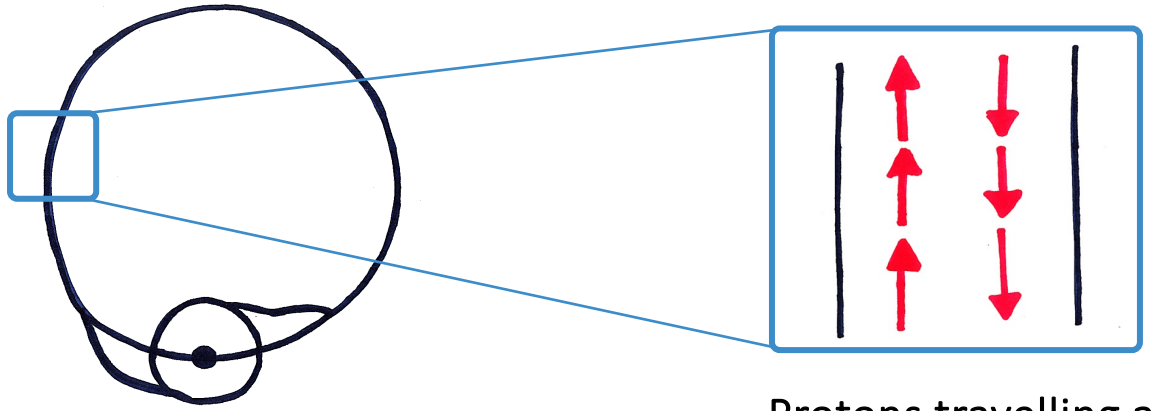
ATLAS EXPERIMENT

Particle Physics Masterclass

Network Workshop

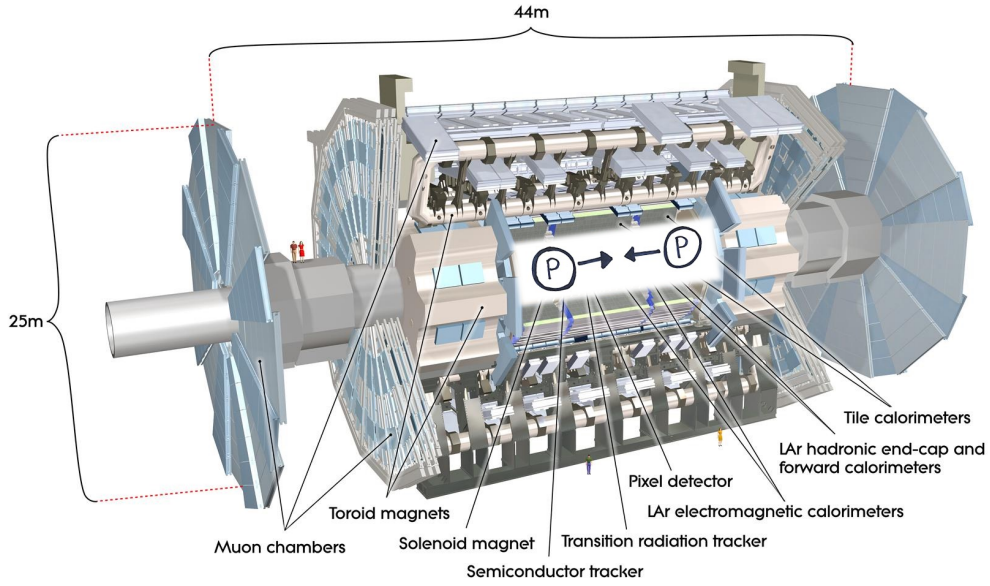
6th September 2024

The ATLAS Experiment at CERN

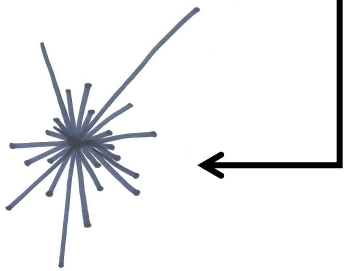


Large Hadron Collider @ CERN

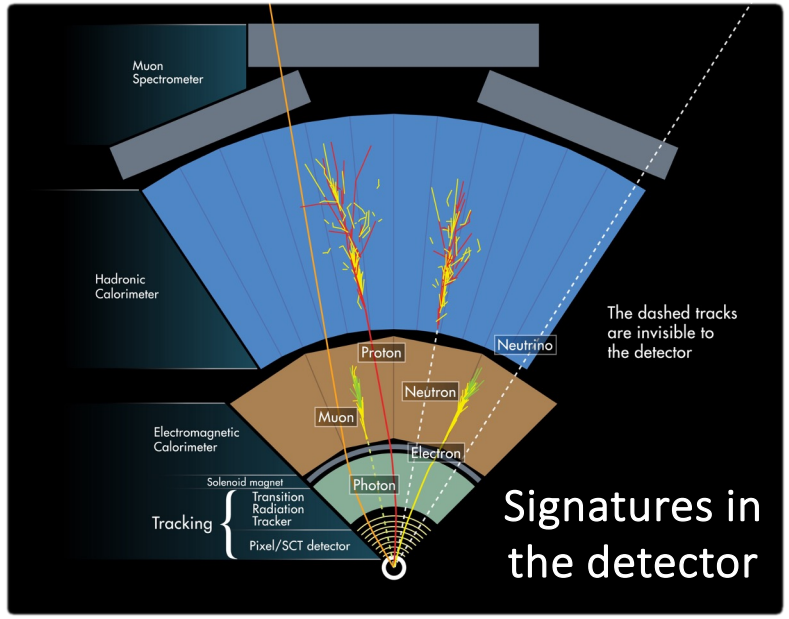
Protons travelling at almost speed of light



Protons collide inside detector (e.g. ATLAS experiment)



Different particles are created in the collisions



Signatures in the detector

Data collection of collision events and respective particles

ATLAS Open Data Values

Accessible

- Make the data and tools package accessible to everyone, keeping in mind the range of internet bandwidths, computer OSs, Mobile access, memory and RAM, and access to experts.

Transferable skills

- Along with particle physics analysis and ATLAS experimental learning objectives, provide skills in programming, software and machine learning

Usable and versatile

- Ensure that many different target audiences, with different backgrounds and skills are able to use the data and tools for a wide range of learning objectives and project goals

Impactful and wide reach

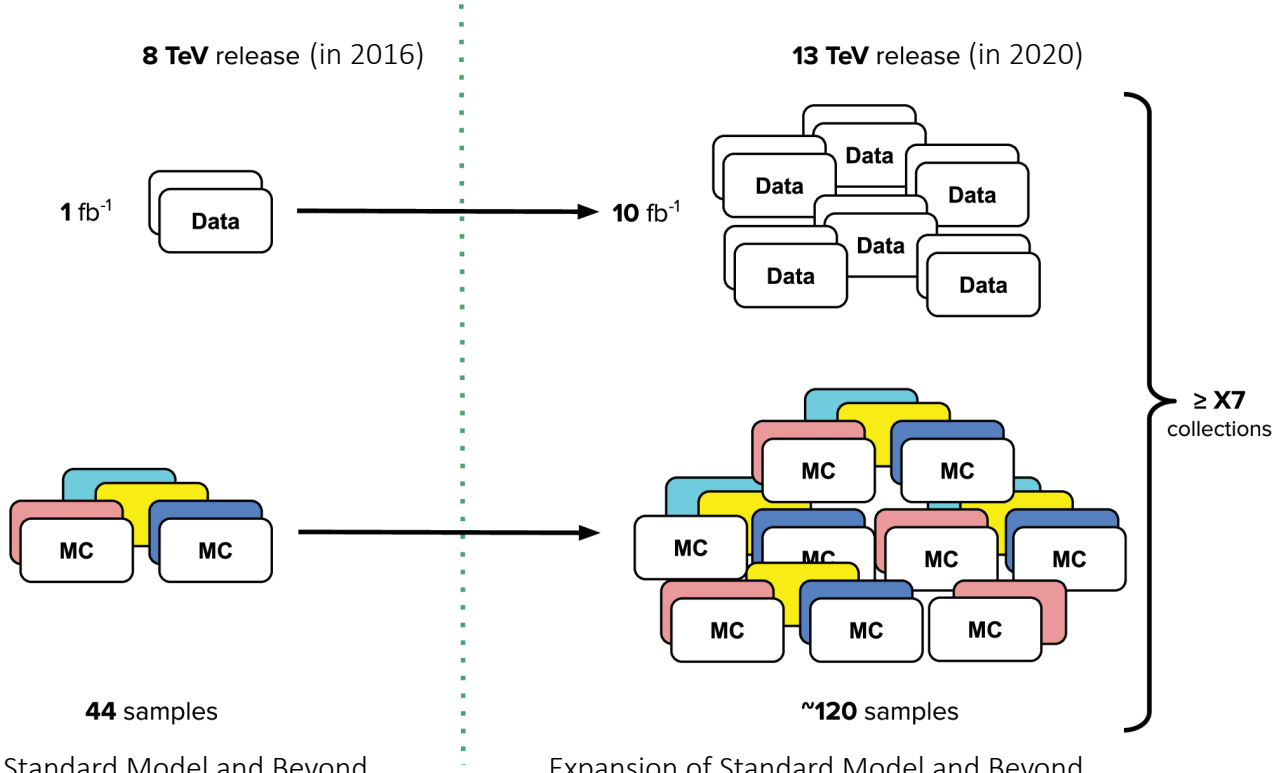
- Push communication campaigns and prepare and deliver online and in person events

Releases Overview

- The **ATLAS experiment collected proton-proton collision data** at centre-of-mass energies of **8 and 13 TeV**, corresponding respectively to 20.2 fb^{-1} (in 2012) and 140 fb^{-1} (2015-2018) of data – **petabytes of raw data**

- Open Data releases address **CERN and ATLAS Open Data policies** (link, link, link)

- **Two public releases of data and Monte Carlo simulated data samples** were released by the ATLAS experiment so far



- Wide range of **physics analysis tools, tutorial videos and data visualisers** provided with the data
- **Datasets, associated information, tools and interactive material accessible via** http://opendata.atlas.cern/
- Extensive dataset information in: Public Note

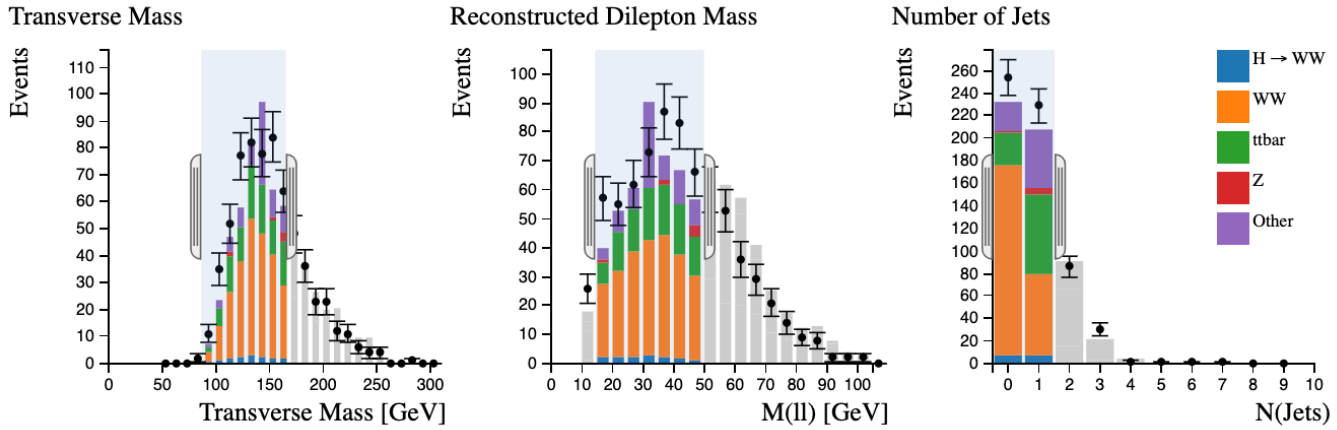
ATLAS Open Data website

- **ATLAS Open Data website** available at opendata.atlas.cern:

ATLAS Open Data Get Started Tutorials Documentation FAQs What's New Contact us

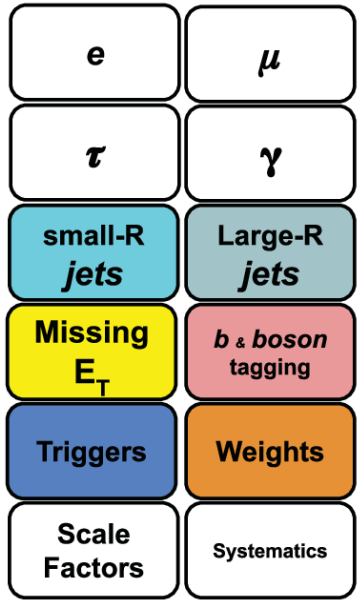
  - Introduction to particle physics and ATLAS detector
  - 8 and 13 TeV datasets and respective documentation
  - Physics analysis and video tutorials for different analyses and methods (ranked by difficulty and time required)
  - Online data visualisation tools
  - Infrastructure: virtual machines, cloud services
  - Contact section

- **Online histogram analyser**
  - Basic entry-level
  - Direct hands-on experience
  - Documentation to provide context and guide students
  - No coding required



Datasets: Current status

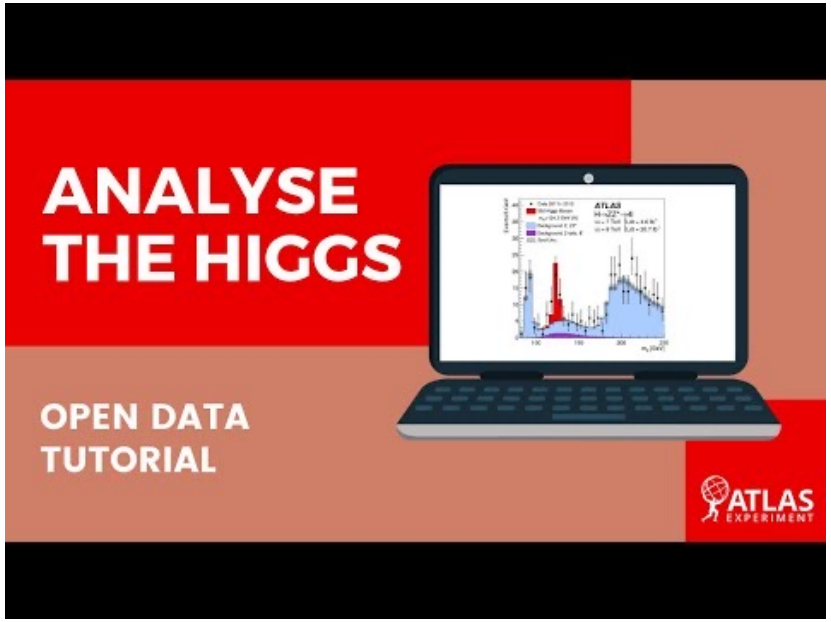
- Datasets contain **simplified information** with respect to the internal ATLAS collected and simulated data
- **Facilitated understanding** and addressing education goal
- **Different datasets/selections on physics objects** allow coverage of **simple to more complex uses**



Around 90 variables

Analysis Tools: Current status

- **Analysis Jupyter n**otebooks are provided for **different analyses and programming languages** (C++, PyROOT, Uproot)
- Notebooks **accessible via** website, along with respective **description** and **difficulty** associated
- Direct access to code also available via **Github** (notebooks, C++, PyROOT, PyROOT extended, Uproot)
- Additional material collected from institutes and Kaggle also available



Introduction Let's take a current ATLAS Open Data sample and create a histogram:

In order to activate the interactive visualisation of the histogram that is later created we can use the JSROOT magic:

```
[ ]: %jsroot on
```

We need to include some standard C++ and ROOT libraries

```
[ ]: // Creates a TChain to be used by the Analysis.C class
#include <TChain.h>
#include <vector>
#include <TFile.h>
#include <iostream>
#include <string>
#include <stdio.h>
```

Because we would like to use more than one ROOT input file, the best option is to use a TChain object. This allows to "chain" several samples into a single structure that we can later loop over

```
[ ]: TString path = "https://atlas-opendata.web.cern.ch/atlas-opendata/samples/2020/GamGam/"
```

```
[ ]: TChain* fChain = new TChain("mini");

fChain->AddFile(path+"Data/data_A.GamGam.root");
fChain->AddFile(path+"Data/data_B.GamGam.root");
fChain->AddFile(path+"Data/data_C.GamGam.root");
fChain->AddFile(path+"Data/data_D.GamGam.root");
```

Now we're going to extract the photons variables

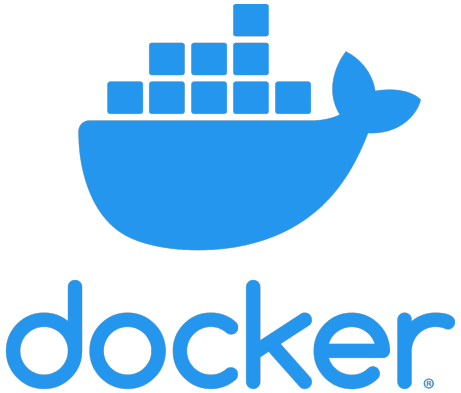
```
[ ]: UInt_t  Photon_n = -1;  //number of preselected photons

vector<float>   *Photon_pt;  //transverse momentum of the photon
```

Software Environment: Current Status

- Big diversity of approaches and possible target audiences
  - Different environments, in which to use analysis code, provided to suit different needs

Online services

Offline services



Usage cases (just a few examples)

- **ATLAS Open Data** is currently **being used** in **schools, universities,** in **public events** and by **interested individuals**

ATLAS Open Data at University of Cape Town, South Africa

- Undergraduate labs at the third-year level
- Major project for third-year undergraduate level
- Minor and major projects in two advanced elective modules

IRIS-UK Project (program), (notebooks)

- More than 300 students from UK schools enrolled, data analysis notebooks developed, posters presented

Internships (e.g. in Technische Universität Dresden, Germany)**, summer programmes** (e.g. in Duke University, North Carolina)

Physics without Frontiers (link)

- **Training** in **Venezuela, Argentina, Uruguay, Honduras, Peru** and many others in Latin America
- Online course in **particle physics** and **machine learning** at the **Royal University of Bhutan**

Kaggle challenges (link)

New release: Datasets

- The ATLAS experiment is adopting a new data format, PHYSLITE, containing reconstructed physics objects already calibrated

- Open Data for Research release
  - Complete PHYSLITE collected and simulated data samples, with **all physics objects** and **systematic uncertainties** (in simulated samples)
  - Respective **documentation provided**
  - Allows for **reproduction of official ATLAS results** and enables **journal-quality research**

- Open Data for Education release
  - **Simplified data format** obtained by filtering/selecting PHYSLITE samples for the most relevant physics objects and systematic uncertainties (in simulated data)
  - **Documentation** is a key part of this effort, particularly given the target audiences

New release: Analysis Tools

- What people want to learn/teach
  - Particle physics
  - Programming (C++ or python)
  - Statistical inferring
  - Machine learning techniques
  - Detector effect corrections
  - Data-driven methods

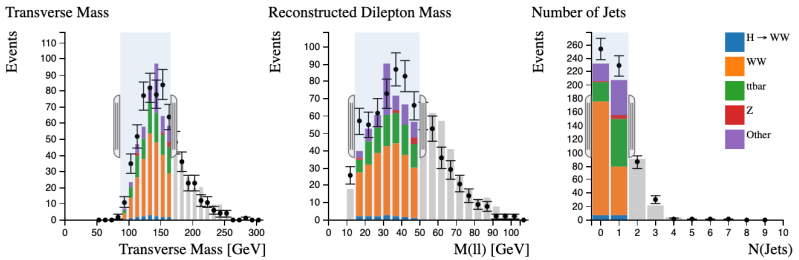
- First step in new release is to **expand current repository of notebooks**
  - Add new/improved analyses and potential new tools
  - Showcase possible physics analyses/final states
  - Different analysis approaches, both in terms of selections and tools
  - Development in Python, C++, etc, depending on personpower and needs
- If resources available, later development of school and university targeted material

Thoughts on the future

Input and potential collaboration very appreciated

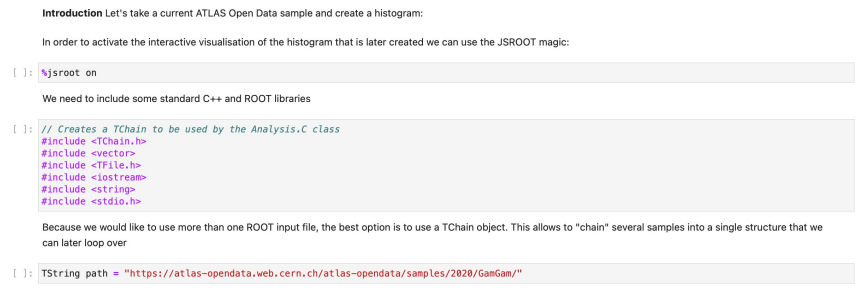
- As before, material can be used in:
  - Summer internships/projects
  - University labs
  - Public events

- Online histogram analysers



- Short activities, well defined walkthrough

- Analysis notebooks



- Medium length activities, insight into analysis

- **Can be used together** with good planning
- **Younger audiences** will require **more basic/direct tools:** either histogram analysers and/or simplified notebooks - **growing interest in development of new tools**
- **Additional material** (documentation, videos, etc) **essential** to deliver meaningful experience

Summary

- **ATLAS Open Data initiative** aims to improve **scientific literacy**, as well foster transfer of highly valuable **skills,** while **reducing accessibility barriers**
- **Two releases** (of proton-proton collisions at 8 and 13 TeV) **already provided to the general public**
- **ATLAS Open Data** website is the starting point to anyone interested in using the data, and where all material can be found
- **Experimental and simulated data** are released to the public
- **Analysis tools, tutorial videos, data visualisers, software and documentation** are provided in addition
- Currently **being used** in **schools, universities,** in **public events** and by **interested individuals**

- Release of **increased 13 TeV dataset** being planned
  - **Documentation being expanded and improved**
  - **New/improved physics analyses and methods** to be covered in **different programming languages**
  - **Open Data for Research** samples allow for **replication** of official results and **journal-quality research**
  - **Open Data for Education** targeting **school/university/public events** applications
  - Goal of reaching even more people with provided material and software environment options

- Please feel free to get in touch: antonio.jacques.costa@cern.ch and kate.shaw@cern.ch

Backup

Public note

Contents

1 Introduction 3

2 Overview of 13 TeV ATLAS Open Data 3

3 13 TeV ATLAS Open Data physics analysis examples 6
- 3.1 Single-lepton final state: the case of SM W-boson production 8
- 3.2 Single-lepton final state: the case of t-channel single-top-quark production 8
- 3.3 Single-lepton final state: the case of top-quark pair production 10
- 3.4 Two-lepton final state: the case of SM Z-boson production 13
- 3.5 Two-lepton final state: the case of SM Higgs boson production in the $H \to WW^*$ decay channel 13
- 3.6 Two-lepton final state: the case of a search for supersymmetric particles 16
- 3.7 Three-lepton final state: the case of SM $W^{\pm}Z$ diboson production 16
- 3.8 Four-lepton final state: the case of SM ZZ diboson production 19
- 3.9 Four-lepton final state: the case of SM Higgs boson production in the $H \to ZZ^*$ decay channel 19
- 3.10 Two-τ-lepton final state: the case of SM Z-boson production 22
- 3.11 Single-lepton boosted final state: the case of a search for BSM $Z' \to t\bar{t}$ 23
- 3.12 Two-photon final state: the case of SM Higgs boson production in the $H \to \gamma\gamma$ decay channel 24

4 General capabilities and limitations of the released 13 TeV dataset 27

5 ATLAS Open Data educational tools 28

6 Summary 28

Appendices 34

A Content of the 13 TeV ATLAS Open Data tuple 34

B MC samples released in the 13 TeV ATLAS Open Data 36

C Evolution of the ATLAS Open Data from the 8 TeV release (2016) to the 13 TeV release (2019) 37

Size, format and contents of datasets

Motivation and validation of physics analyses

Details on released tools

Datasets: Current status

- Datasets contain **simplified information** with respect to the internal ATLAS collected and simulated data
- Facilitated understanding and addressing education goal
- Datasets made **available with different selections on physics objects** (leptons, photons, etc)
- Different datasets/selections allow coverage of **simple to more complex uses**
- Data is stored in a “flat ntuple format”, and provides the physics objects reconstructed at the detector level, along with experimental calibrations/corrections associated

Tuple branch name	C++ type	Variable description
runNumber	int	number uniquely identifying ATLAS data-taking run
eventNumber	int	event number and run number combined uniquely identifies event
channelNumber	int	number uniquely identifying ATLAS simulated dataset
mcWeight	float	weight of a simulated event
XSection	float	total cross-section, including filter efficiency and higher-order correction factor
SumWeights	float	generated sum of weights for MC process
scaleFactor_PILEUP	float	scale-factor for pileup reweighting
scaleFactor_ELE	float	scale-factor for electron efficiency
scaleFactor_MUON	float	scale-factor for muon efficiency
scaleFactor_PHOTON	float	scale-factor for photon efficiency
scaleFactor_TAU	float	scale-factor for tau efficiency
scaleFactor_BTAG	float	scale-factor for b-tagging algorithm @70% efficiency
scaleFactor_LepTRIGGER	float	scale-factor for lepton triggers
scaleFactor_PhotonTRIGGER	float	scale-factor for photon triggers
trigE	bool	boolean whether event passes a single-electron trigger
trigM	bool	boolean whether event passes a single-muon trigger
trigP	bool	boolean whether event passes a diphoton trigger
lep_n	int	number of pre-selected leptons
lep_truthMatched	vector<bool>	boolean indicating whether the lepton is matched to a simulated lepton
lep_trigMatched	vector<bool>	boolean indicating whether the lepton is the one triggering the event
lep_pt	vector<float>	transverse momentum of the lepton
lep_eta	vector<float>	pseudo-rapidity, η, of the lepton
lep_phi	vector<float>	azimuthal angle, ϕ, of the lepton
lep_E	vector<float>	energy of the lepton
lep_z0	vector<float>	z-coordinate of the track associated to the lepton wrt. primary vertex
lep_charge	vector<int>	charge of the lepton
lep_type	vector<int>	number signifying the lepton type (e or μ)
lep_isTightID	vector<bool>	boolean indicating whether lepton satisfies tight ID reconstruction criteria
lep_ptcone30	vector<float>	scalar sum of track p_{T} in a cone of R=0.3 around lepton, used for tracking isolation
lep_etcone20	vector<float>	scalar sum of track E_{T} in a cone of R=0.2 around lepton, used for calorimeter isolation
lep_trackd0pvunbiased	vector<float>	d_0 of track associated to lepton at point of closest approach (p.c.a.)
lep_tracksigd0pvunbiased	vector<float>	d_0 significance of the track associated to lepton at the p.c.a.
met_et	float	transverse energy of the missing momentum vector
met_phi	float	azimuthal angle of the missing momentum vector
jet_n	int	number of pre-selected jets
jet_pt	vector<float>	transverse momentum of the jet
jet_eta	vector<float>	pseudo-rapidity, η, of the jet
jet_phi	vector<float>	azimuthal angle, ϕ, of the jet
jet_E	vector<float>	energy of the jet
jet_jvt	vector<float>	jet vertex tagger discriminant [21] of the jet
jet_trueflav	vector<int>	flavour of the simulated jet
jet_truthMatched	vector<bool>	boolean indicating whether the jet is matched to a simulated jet
jet_MV2c10	vector<float>	output from the multivariate b-tagging algorithm [22] of the jet

Analysis Tools: Current status

- **Analysis Jupyter n**otebooks are provided for **different analyses and programming languages** (C++, PyROOT, Uproot)
- Notebooks **accessible via** website, along with respective **description** and **difficulty** associated
- Direct access to code also available via **Github:**
  - https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata
  - https://github.com/atlas-outreach-data-tools/atlas-outreach-cpp-framework-13tev/tree/master
  - https://github.com/atlas-outreach-data-tools/atlas-outreach-data-tools-framework/tree/master/Analysis
  - https://github.com/atlas-outreach-data-tools/atlas-outreach-Python-uproot-framework-13tev
  - https://github.com/atlas-outreach-data-tools/atlas-outreach-PyROOT-framework-13tev
- Additional material collected from institutes and Kaggle also available